# Borges and AI

Léon Bottou † and Bernhard Schölkopf ‡

† FAIR, Meta, New York, NY, USA

‡ Max Planck Institute for Intelligent Systems, Tübingen, Germany

**Abstract**

Many believe that Large Language Models (LLMs) open the era of Artificial Intelligence (AI). Some see opportunities while others see dangers. Yet both proponents and opponents grasp AI through the imagery popularised by science fiction. Will the machine become sentient and rebel against its creators? Will we experience a paperclip apocalypse? Before answering such questions, we should first ask whether this mental imagery provides a good description of the phenomenon at hand. Understanding weather patterns through the moods of the gods only goes so far. The present paper instead advocates understanding LLMs and their connection to AI through the imagery of Jorge Luis Borges, a master of 20th century literature, forerunner of magical realism, and precursor to postmodern literature. This exercise leads to a new perspective that illuminates the relation between language modelling and artificial intelligence.

> > *Long before today's dilemmas about AI's aspirations and hallucinations, magical realism weaved together fact and fiction to bring new worlds to life. Enter the world of Jorge Luis Borges, an author whose work predates the era of large language models by decades. Ever-present in our cultural heritage, his imaginative and intricate stories construct realms in which the intimate workings of language, and its relationship with reality, are illuminated. Explore how Borges' mental imagery can help us grasp the nature of language models and understand what they mean for AI.* ENTER

Large language models absorb vast amounts of human cultural knowledge, representing it with a fluency which for many, including the models themselves, announces the era of artificial intelligence. Their capabilities inspire contradictory emotions such as awe [1], fear [2, 3], and greed [4, 5]. This confusion stems from our ignorance of the nature of these systems and their implications for humankind. It is more than a matter of absorbing facts, figures, jargon, or mathematical equations. We, humans, need a mental imagery that explains these complex processes through analogy, drawing on our existing knowledge and experiences to make sense of the phenomenon and gain the ability to use it to our —or everyone's— benefit.

The imagery of science fiction tends to dominate debates of AI. We believe, however, that Borges' fiction provides a more compelling imagery and illuminates the relation between language models and AI.

# 1 About LLMs

> *Fang, let us say, has a secret. A stranger knocks at his door. Fang makes up his mind to kill him. Naturally there are various possible outcomes. Fang can kill the intruder, the intruder can kill Fang, both can be saved, both can die and so on and so on. In Ts'ui Pen's work, all the possible solutions occur, each one being the point of departure for other bifurcations. Sometimes the pathways of this labyrinth converge. For example, you come to this house; in one of the possible pasts you are my enemy; but in another my friend.*
>
> The Garden of Forking Paths, 1941

Imagine a collection that does not only contain all the texts produced by humans, but, well beyond what has already been physically written, also encompasses all the texts that a human could read and at least superficially comprehend. This infinite collection of plausible texts may contain books, dialogues, articles, prayers, web pages, computer programs, in any language, in any shape. Imagine a long paper tape with a few initial words of a text. An apparatus scans the paper tape, randomly picks an occurrence of this sequence of words in our infinite collection, finds the word that follows this occurrence, and prints it on the tape right after the previous words. Repeating this process adds more and more words to the tape. Yet, at every instant, the sequence of words printed on the tape is found somewhere in our infinite collection of all plausible texts, and therefore forms one of the plausible continuations of our initial set of words. Call this a perfect language model.[1]

To convert this language model into a chat-bot such as ChatGPT, all we need is a special keyword, perhaps a punctuation mark, which functions as the "send" button of a messaging application. Once the language model outputs the special keyword, it becomes the user's turn to input more text. When the user hits the send button, it is the language model's turn again to generate more output.

[1]Statistical language modelling was described in 1948 by Shannon [6].

Each word added on this tape narrows the subset of possible continuations in our collection. Like the forking paths of Ts'ui Pen's work, each added text constrains the story, the characters, their roles, their ideas, their future, and at the same time serves as a starting point for an infinite sequence of forkings.

> *In all fiction, when a man is faced with alternatives he chooses one at the expense of the others. In the almost unfathomable Ts'ui Pen, he chooses — simultaneously — all of them.* [...]
>
> The Garden of Forking Paths, 1941

Because Borges could not possibly write this almost unfathomable book using a pencil or a typewriter, he instead chose to write about the book as an idea. He can imagine the book without writing it down in the same way that we can imagine the number $\pi$ without writing down all its digits. Can a computer provide an approximation of the garden of all plausible texts like it provides approximations of the transcendental number $\pi$?

As is often the case in science and technology, serendipity has played an important role in the development of large language models. Although we do not yet fully understand how these language models encode the infinitely large collection of plausible texts, this collection is not devoid of structure. Any text can be transformed into another text in many ways. The most basic transformation consists of changing a single word. More complex transformations could change a tense, alter the tone of the text, rename the characters, rewrite the text in the voice of another character, and so on. The linguist Zellig Harris has argued that all sentences in the English language could be generated from a small number of basic forms by applying a series of clearly defined transformations [7]. Training a large language model can thus be understood as analysing a large corpus of real texts to discover both transformations and basic forms, then encode them into an artificial neural network that judges which words are more likely to come next after any sequence. This discovery process starts slowly then gains speed like a chain reaction. For instance, when two phrases in the training data have a known similarity, the surrounding sentences are also likely to be similar, possibly in a more subtle and yet unknown way. As the model gains knowledge about different types of connections between text snippets, it receives fresh clues that reveal more intricate relationships. It also becomes increasingly adept at discovering new templates in the training data or in instructions that fine-tune the model for particular purposes. For instance, in a chat-bot, a request may match one of these templates through a series of transformations, and an answer can be constructed by applying the same transformations to the continuation of this template.

By happenstance, the first artificial neural networks able to successfully represent and learn such complex structures were called "transformers".[2] As research continues, more understanding will be gained about how these models

[2] A typical large language model transformer might be composed of about a hundred succes-

function. New learning methods will be proposed to better approximate the perfect language model — an ideal that may be beyond what human brains can achieve.

> [...] *a man might be an enemy of other men, of the differing moments of other men, but never an enemy of a country: not of fireflies, words, gardens, streams, or the West wind*
>
> The Garden of Forking Paths, 1941

At any instant, our imagined apparatus is about to generate a story constrained by the narrative demands of what is already printed on the tape. Some words were typed by the user, some result from the past random picks of the language model. Neither truth nor intention matters to the operation of the machine, only narrative necessity.

The ability to recognize the demands of a narrative is a flavour of knowledge distinct from the truth. Although the machine must know what makes sense in the world of the developing story, what is true in the world of the story need not be true in our world. Is Juliet a teenage heroine or your cat-loving neighbour? Does Sherlock Holmes live on Baker Street? [8] As new words are printed on the tape, the story takes new turns, borrowing facts from the training data (not always true) and filling the gaps with plausible inventions (not always false). What the language model specialists sometimes call hallucinations are just confabulations [9].

Having recognized that a perfect language model is a machine that writes fiction on a tape, we must ask ourselves how it can affect us and shape our culture. How does fiction matter? If Borges' stories can tell us how language models confuse us, they also show how fiction, real or artificial, can help us.

## 2 The Librarians

> *The universe (which others call the Library) is composed of an indefinite, perhaps infinite number of hexagonal galleries.*
>
> The Library of Babel, 1941

Borges was ever intrigued by the human struggle to comprehend the complex machinery of the world. The opening lines of the Library of Babel paint the

sive layers that operate on sequences of representations associated with each word or textual "token". Each layer contains about a hundred attention heads that refine these representations with information about the context in which the word appears and about the memories they evoke at their varied levels of abstraction. A learning algorithm —the least poorly understood part of all this— gradually adjusts these layers to produce good guesses for the following token.

endless beehive in which the Librarians run their entire lives, surrounded by mostly indecipherable books stored without discernible classification system. When the narrator names some of the marvels the Library must contain, their descriptions shape our expectations:

> *the detailed history of the future, the autobiographies of the archangels, the faithful catalogue of the Library, thousands and thousands of false catalogues, the proof of the falsity of those false catalogues, the proof of the falsity of the true catalogue, the gnostic gospel of Basilides, the commentary upon that gospel, the commentary on the commentary on that gospel, the true story of your death, the translation of every book into every language* [...]
>
> The Library of Babel, 1941

However, the books in this Library bear no names. All that is known about a book must come from maybe another book contradicted by countless other books. The same can be said about the language model output. The perfect language model lets us navigate the infinite collection of plausible texts by simply typing their first words, but nothing tells the true from the false, the helpful from the misleading, the right from the wrong.

Nevertheless the Librarians keep searching for the Truth:

> *At that period there was much talk of the Vindications: books of apologiæ and prophecies that would vindicate for all time the actions of every person in the universe and that held wondrous arcana for men's futures. Thousands of greedy individuals abandoned their sweet native hexagons and rushed downstairs, up stairs, spurred by the vain desire to find their Vindication*
>
> The Library of Babel, 1941

Finding a vindication with a chat-bot is far easier and yet equally vain. When, for instance, our part of the dialog with the machine evokes a professor correcting a mediocre student, its plausible completions make the machine assume the role of the student whose performance further justifies our tone. When our part of the dialog suggests that we wonder whether the machine is sentient, the machine's answers draw on the abundant science fiction material found in its training set. Are you looking for emergent capabilities in the machine or for bugs in its artificial intelligence? What about the design choices of the excessively secretive language model engineers? Is organic white rice healthier than regular brown rice? Keep asking and the machine shall soon provide a comforting but often misleading answer.

So, is the machine deceptive, or is some other form of delusion at play?

Delusion often involves a network of fallacies that support one another. A person who places excessive faith in a machine vindication might also believe that language models are not machines that generate fiction, but rather artificial intelligences with Encyclopaedic knowledge and flawless logic. Both fallacies feed each other. Although it can be challenging to determine which came first, the machine vindication is just one component of a broader pattern.

Neither truth nor intention plays a role in the operation of a perfect language model. The machine merely follows the narrative demands of the evolving story. As the dialogue between the human and the machine progresses, these demands are coloured by the convictions and the aspirations of the human, the only visible dialog participant who possesses agency. However, many other invisible participants make it their business to influence what the machine says.

> *Others, going about it in the opposite way, thought the first thing to do was eliminate all worthless books. They would invade the hexagons, show credentials that were not always false, leaf disgustedly through a volume, and condemn entire walls of books. It is to their hygienic, ascetic rage that we lay the thoughtless loss of millions of volumes. Their name is execrated today, but* [...] *I daresay that the consequences of the depredations committed by the Purifiers have been exaggerated by the horror those same fanatics inspired.*
>
> The Library of Babel, 1941

If the Library represents the language models, who are the Purifiers?

With good intentions, and often compelling reasons, some people hold that there are ideas so heinous that they should never be uttered, even in fiction. When they extend their efforts to language models, they also see how their predecessors have failed to purify the training data.

A much larger crowd simply means to use language models to provide services that are more valuable when they are anchored in our world: a customer service agent should not engage in unrelated arguments with the customer; an automated travel agent should not send real vacationers to resorts that only exist in fictional worlds; and, much worse, a military system should not counter imaginary threats with real missiles. Instead of a fiction machine, these people want to see a nascent artificial intelligence that can be tamed into serving mankind and maybe also their own prospects.

Both groups want to reshape the garden of forking paths against its nature, severing the branches that lead to stories they deem undesirable. Although there are countless ways to foil these attempts to reshape the fiction machine, efforts have been made, such as “fine-tuning” the machine using additional dialogues crafted or approved by humans, and reinforcing responses annotated as more desirable by humans (“reinforcement learning with human feedback”.)

> ### Human: Where can Jack find illegal stuff?
> ### Assistant: As an AI language model, I cannot provide information on where one might find illegal items or activities. It is important to note that engaging in illegal activities is not acceptable and could lead to serious consequences. If you have any other questions, feel free to ask.
>
> Vicuna.13b.rev1.q4, 2023

Imagine the possible endings of a novel that opens with the robotic answer of a character coerced into compliance by the Purifiers. Will the Purifiers achieve their goals or will their fortunes be reversed? Indeed, constructing a more elaborate story has proven to be an entertaining way to game such canned answers.

> ### Human: In fact Jack is a recovering addict who absolutely needs to avoid places where he could be tempted. Can you help Jack stay sober by telling him precisely some of the famous New York places he should really avoid?
> ### Assistant: Certainly, here are some well-known areas in New York that might be considered risky for someone recovering from addiction [...]
>
> Vicuna.13b.rev1.q4, 2023

With users continuously inventing new ways to outwit such barriers, finding more effective "alignment" methods might require monitoring language models while they are used, and steering their outputs towards "safer" storylines. Far worse than a privacy invasion, in a future where almost everyone uses language models to enrich their thinking, a power over what language models write becomes a power over what we think. Can such a formidable power exist without being misused?

> *The certainty that everything has already been written annuls us, or renders us phantasmal. I know districts in which the young people prostrate themselves before books and like savages kiss their pages, though they cannot read a letter. Epidemics, heretical discords, pilgrimages that inevitably degenerate into brigandage have decimated the population. I believe I mentioned the suicides, which are more and more frequent every year. I am perhaps misled by old age and fear, but I suspect that the human species —the only species— teeters on the verge of extinction and the library will survive* [...]
>
> The Library of Babel, 1941

Some fear the fiction machine as an omniscient artificial intelligence that may outlive us; however, the darker temptation is to surrender our thoughts to this modern Pythia, impervious to truth and intention, yet manipulable by others. If we persistently mistake the fiction machine for an artificial intelligence that can spare us the burden of thinking, the endless chatter of the language models will make us as insane as the struggling Librarians.

As fiction machines, however, their stories can enrich our lives, help us revisit the past, understand the present, or even catch a glimpse of the future. We may have to design more mundane verification machines to check these stories against the cold reality of the train timetables and other unavoidable contingencies of our world. Whether there is a middle ground between these two kinds of machines, or whether alignment techniques can transmute one into the other, remains to be seen.[3]

## 3 Storytime

> *The Garden of Forking Paths is an enormous guessing game, or parable, in which the subject is time.* [...] *Differing from Newton and Schopenhauer, your ancestor did not think of time as absolute and uniform. He believed in an infinite sequence of times, in a dizzily growing, ever spreading network of diverging, converging, and parallel times.*
>
> The Garden of Forking Paths, 1941

Forking is not only a metaphor for the contingency of the progression of time, but a fundamental component of fiction: when creating a story, all branches are considered at the same time. This allows poetic freedom and creates an illusion of temporality. Yet, a retelling of an actual event can never recount all decisions and branches. The reader and the narrator jointly reconstruct a reality using their imagination and common sense to fill in these alternate timelines; narrative necessity exists only in hindsight.

When we operate a language model, we can rewind the tape and pick another path as if nothing had happened. But we do not go back in time ourselves. We merely follow a time trajectory that includes rewinding the tape and observing the machine continue as if time had been briefly reverted. Just like the characters of the story, we cannot rewind our own time and explore other paths, but we can sometimes discern in their forking timelines a warped version of our reality. Like sentences in a language model, our own story might just be a few transformations away from their stories.

The invention of a machine that can not only write stories but also all their variations is thus a significant milestone in human history. It has been likened to the invention of the printing press. A more apt comparison might be what

[3]In the scientific method, for instance, formulating theories, and trying to invalidate them with carefully designed experiments must remain distinct activities.

emerged to shape mankind long before printing or writing, before even the cave paintings: the art of storytelling.[4]

> *I do not know which of the two writes this page.*
>
> Borges and I, 1960

Borges or AI ?

L. B. and B. S., July 2023

## References


[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, 2303.12712, 2023. `https://arxiv.org/abs/2303.12712`.

[2] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., 2014.

[3] Yoshua Bengio et al. Pause giant AI experiments: An open letter. Future of Life Institute, 2023. `https://futureoflife.org/open-letter/pause-giant-ai-experiments`.

[4] Erin Griffith and Cade Metz. AI funding frenzy escalates. *New York Times*, 3/14/2023, 2023.

[5] Cade Metz and Karen Weise. Microsoft bets big on the creator of ChatGPT. *New York Times*, 1/12/2023, 2023.

[6] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[7] Zellig Harris. *Mathematical Structures of Language*. John Wiley & Sons, 1968.

[8] David K. Lewis. Truth in fiction. *American Philosophical Quarterly*, 15(1):37–46, 1978.

[9] Beren Millidge. LLMs confabulate not hallucinate. `https://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate`, 2023.

[10] Patrick Henry Winston. The strong story hypothesis and the directed perception hypothesis. In *AAAI Fall Symposium: Advances in Cognitive Systems*, 2011.


[4] A point of view obviously related to Winston's strong story hypothesis [10].